Automatic Classification of HCV and HIV-1 Sequences with the Branching Index

Peter Hraber, Carla Kuiken, Mark Waugh, Shaun Geer, William J. Bruno,
Thomas Leitner, T-10

Knowledge of viral subtype is important for molecular epidemiology, for identification of new variants and circulating recombinant forms, and for informed treatment options. Viral sequences are tremendously variable, presenting a challenge for systematic classification.

Automatic classification of viral sequences should be fast, objective, accurate, and reproducible. Most methods that classify sequences use either pairwise distances or phylogenetic relations. However, they cannot discern when a sequence is unclassifiable. The branching index (BI) combines distance and phylogeny methods to compute a ratio that quantifies how closely a query sequence clusters with a subtype clade (Fig. 1).

In the hypothesis-testing framework of statistical inference, the BI is compared with a threshold to test whether sufficient evidence exists for the query sequence to be classified among known sequences. If above the threshold, the null hypothesis of no support for the subtype relation is rejected, and the sequence is taken as belonging to the subtype clade with which it clusters on the tree.

We studied statistical properties of the BI for subtype classification in HCV and HIV-1. Pairs of BI values with known positive and negative test results (Fig. 2) were computed from 10,000 random fragments of reference alignments. Sampled fragments were of sufficient length to contain phylogenetic signal that properly groups reference sequences together into subtype clades. For HCV, a threshold BI of 0.71 yields 95.1% agreement with reference subtypes, with equal false positive and false negative rates. For HIV-1, a threshold of 0.66 yields 93.5% agreement. Higher thresholds can be used where lower false positive rates are required.

BI profiles result from moving windows over the length of the query sequence. Shown graphically, BI profiles enable visualization of classification consistency over the query sequence (Fig. 3). A new service for automatic classification of HIV-1 and HCV sequences with the branching index is being provided online.

This classification approach might also be used to classify influenza and other viral genotypes, to identify bacterial subspecies, for molecular subtyping of bacterial toxins, and for more general use in clustering algorithms and supervised learning.

For further information contact Peter Hraber at phraber@lanl.gov.

Funding Acknowledgements
- National Institutes of Health
- Department of Energy

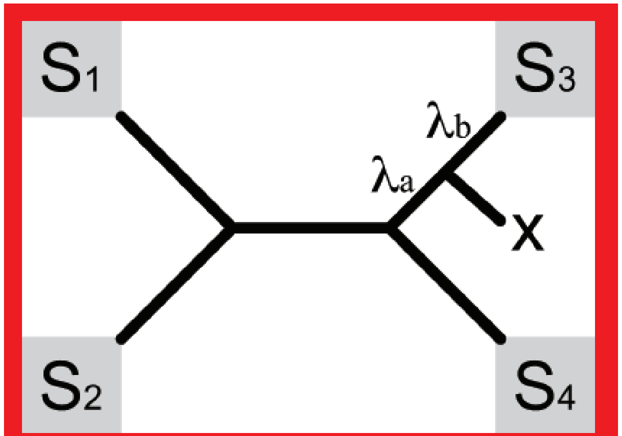

Fig. 1. BI illustration. The branching index is defined for query sequence x and subtype clades S_{1-4} as the ratio of branch lengths $\lambda_a / (\lambda_a + \lambda_b)$, given proximity to S_3.

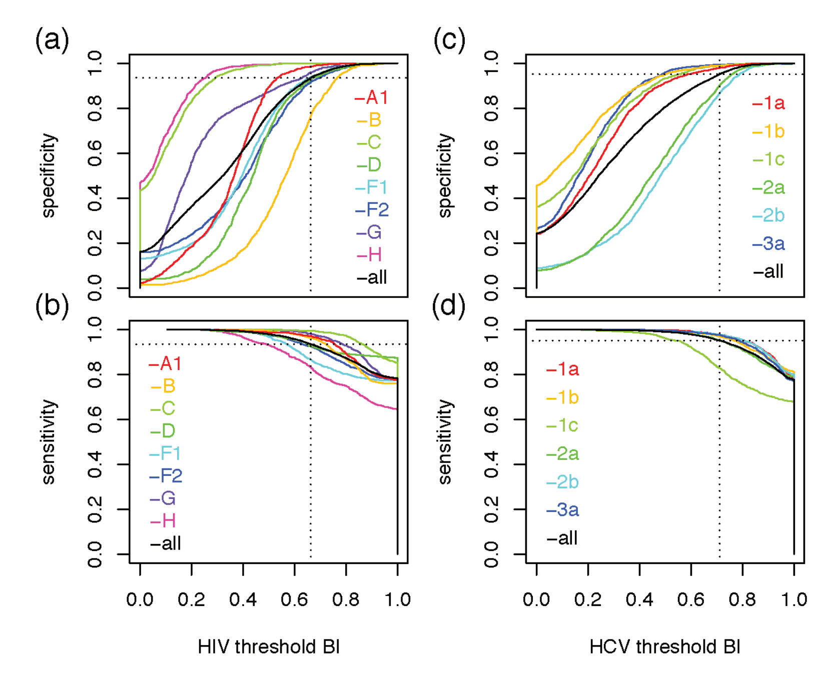

Fig. 2. Error rates from tests with known outcomes. Cumulative distributions from 10,000 random HIV-1 fragments show specificity (a) and sensitivity (b) for any threshold, and likewise for HCV specificity (c) and sensitivity (d). Line color indicates viral subtype. Dotted lines indicate thresholds that optimize classification accuracy for all samples.

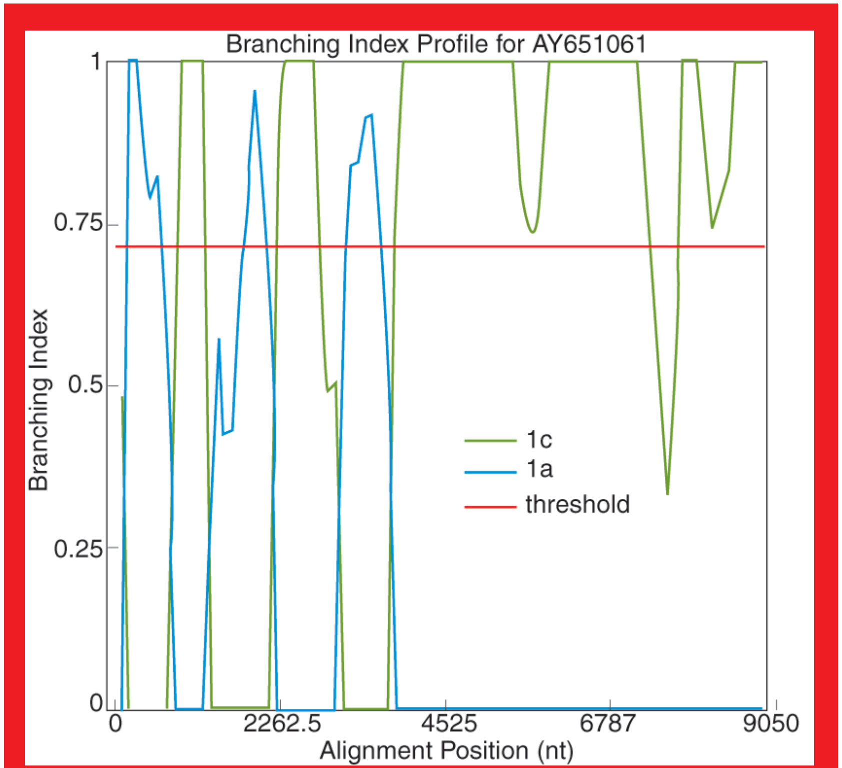

Fig. 3. BI profile from HCV sequence AY651061 (400 nt windows with 40 nt between each successive window) indicates a 1a/1c recombinant.